Empirical Economics (2021) 60:353–363
https://doi.org/10.1007/s00181-020-01929-w

A hierarchical panel data stochastic frontier model for the estimation of stochastic metafrontiers

Christine Amsler[1] · Yi Yi Chen[2] · Peter Schmidt[1] · Hung Jen Wang[3]

Received: 29 January 2020 / Accepted: 6 August 2020 / Published online: 19 August 2020

Abstract
This paper proposes a stochastic frontier model with three composed errors, and therefore six error components. As in the metafrontier literature, firms belong to groups with a group-specific frontier. A firm has a level of short-run and long-run inefficiency relative to its group-specific frontier, as in existing models with two composed errors and four error components. But now there is also a group-specific inefficiency, that is, a shortfall of the group-specific frontier from the best practice metafrontier. The paper shows how to estimate this model and how to extract predictions of the various inefficiencies.

Keywords Stochastic frontier · Panel data · Hierarchical model · Metafrontier · Inefficiency

JEL Classification C23 · C26

1 Introduction

Several recent papers have proposed models to separate long-run and short-run technical inefficiency from each other and from long-run and short-run heterogeneity that is not regarded as inefficiency. In these papers, there are two composed errors, each of which has two parts. There is a long-run (time-invariant) composed error $c_i = c_i^o - c_i^*$, where c_i^o is normal and represents long-run heterogeneity, and $c_i^* \geq 0$ is half-normal and represents long-run inefficiency. There also is a short-run (independent over time) composed error $u_{it} = u_{it}^o - u_{it}^*$, where u_{it}^o is normal and represents short-run het-

✉ Peter Schmidt
schmidtp@msu.edu

1 Michigan State University, East Lansing, USA

2 Tamkang University, New Taipei City, Taiwan

3 National Taiwan University, Taipei City, Taiwan

erogeneity, and $u_{it}^* \geq 0$ is half-normal and represents short-run inefficiency. Papers that have considered models with this error structure include Colombi et al. (2011), Kumbhakar et al. (2014), Colombi et al. (2014) [hereafter CKMV], Tsionas and Kumbhakar (2014), Filippini and Greene (2016) and Lai and Kumbhakar (2018). Despite the rather large number of error components, this model has been estimated successfully. As stressed by Filippini and Greene, this is not so surprising if we view the model as containing two composed errors, as opposed to four error components.

In this paper we will add a third composed error $w_g = w_g^o - w_g^*$, where $g = g(i)$ represents the "group" or "cluster" or "industry" to which firm i belongs. We will use the word "group" but in fact g can be anything that clusters the firms, like industry or geographical region or choice of technology.

The motivation for this model is the concept of a "metafrontier." The idea of a metafrontier originated with Hayami and Ruttan (1971, 1985), and was later operationalized in a large number of other papers, including Pitt (1983), Lau and Yotopoulos (1989), Battese and Rao (2002), Battese et al. (2004), O'Donnell et al. (2008), Moreira and Bravo-Ureta (2010), Villano et al. (2015) and Amsler et al. (2017). In this literature one is interested in the technical inefficiency of a firm relative to its group-specific frontier, but also in the "technology gap" between the group-specific frontier and the overall maximal frontier (the metafrontier). That is, for each group there is a frontier and a firm in that group has an inefficiency relative to that frontier, but there is also a potential inefficiency from being in an inefficient group (e.g. using the wrong technology), reflected in the distance between the group-specific frontier and the metafrontier. In our model this technology gap, for firms in group g, is captured by w_g^*.

In the original work of Hayami and Ruttan, the groups corresponded to different choices of technology for growing rice (traditional methods versus "green revolution" methods), but as noted above the groups can be defined in many different ways in different settings. For example, in Amsler, O'Donnell and Schmidt, the "firms" are actually countries and the "groups" are continents. The only requirement is that the number of groups and the group membership of each firm are known.

2 Model and notation

Suppose that we have T time periods, indexed by $t = 1, \ldots, T$, and G groups, indexed by $g = 1, \ldots, G$. We have n_g firms in group g, so that the total number of firms is $N = \sum_g n_g$. We will index firms by $i = 1, \ldots, N$. We will assume that each firm is in only one group and that firms do not change groups over time, so that we can represent the group to which firm i belongs as $g(i)$.

This is a "hierarchical" or "multi-level" data structure. Firms are nested in groups, where "nested" is a term that dates back to the seminal article of Fuller and Battese (1973), because for each firm there is a unique group. By way of contrast, time is not nested in either firms or groups. There is a very large literature on hierarchical models (i.e. linear models for hierarchical data). A very selective list of references includes Fuller and Battese (1973), Raudenbush and Bryck (2002), Kim and Frees (2007), Wooldridge (2010, pp. 876–883) and Matyas (2017), the latter being a recent comprehensive treatment of the topic.

Our model will be similar to the model of Yang and Schmidt (2020), which has fixed time effects and random firm and group effects. Specifically, our model is:

$$y_{it} = x_{it}^{'}\beta + d_t^{'}\theta + c_i + w_{g(i)} + u_{it}. \tag{1}$$

Here y_{it} is the output (in logs) of firm i at time t; x_{it} is a vector of measures of inputs; and d_t is a dummy variable for time t, so that the elements of θ are the fixed time effects. The c_i, w_g and u_{it} are the long-run, group-specific and short-run composed errors, respectively. Thus $c_i = c_i^o - c_i^*$, where c_i^o is normal and $c_i^* \geq 0$ is half-normal; similarly, $w_g = w_g^o - w_g^*$ and $u_{it} = u_{it}^o - u_{it}^*$. These composed errors have a skew-normal distribution (Azzalini 1985).

There are some minor differences between this model and the model of Yang and Schmidt. For example, their model distinguishes time-varying inputs (their x_{it}) from time-invariant inputs (their w_i), which is a relevant distinction for fixed-effects or generalized least squares estimation but not a relevant distinction in this paper. Also their model contains group-specific variables (their z_g). We do not include these because it is not reasonable to think of group-specific variables as inputs, though they could be included as "environmental variables" in a model for the distribution of w_g^*. But the main difference is that our c_i, w_g and u_{it} are composed errors as opposed to random effects of the usual kind (zero mean and unspecified distribution).

Our interpretation of the model is as follows. The overall frontier (metafrontier) for firm i at time t is $y_{it} = x_{it}^{'}\beta + d_t^{'}\theta + \left(c_i^o + w_{g(i)}^o + u_{it}^o\right)$. Here c_i^o is firm-specific heterogeneity; $w_{g(i)}^o$ captures heterogeneity across groups; and u_{it}^o is idiosyncratic heterogeneity. For firm i at time t, its inefficiency relative to the overall frontier is $\left(c_i^* + w_{g(i)}^* + u_{it}^*\right)$; its inefficiency relative to its group-specific frontier is $c_i^* + u_{it}^*$; and $w_{g(i)}^*$ is the inefficiency of its group relative to the metafrontier. So we have split the inefficiency of firm i at time t relative to the overall frontier $\left(c_i^* + w_{g(i)}^* + u_{it}^*\right)$ into its inefficiency relative to its group-specific frontier $(c_i^* + u_{it}^*)$ plus the inefficiency of group $g(i)$ relative to the best practice frontier $(w_{g(i)}^*)$.

Many applications that we can envision would have a small number of time periods (T), a small number of groups (G), but a large number of firms per group (n_g) and therefore a large total number of firms (N). Our model has fixed time effects, which is consistent with small T and with the fact that we simply want to control for the time effects, not to decompose them into heterogeneity and inefficiency. We have random firm effects, for two reasons. First, if we had fixed firm effects, the group effects would not be identified, due to the nested nature of the data. Second, we are interested in decomposing the firm effects into heterogeneity and inefficiency, and this requires them to be random and to obey our normal/half-normal distributional assumptions. We could have fixed group effects, but once again decomposing them into their heterogeneity and inefficiency components requires random effects and distributional assumptions. A potential issue is that the decomposition of the group effects into their heterogeneity and inefficiency components will require estimates of the variances of w^o and w^*, and these parameters cannot be expected to be estimated very precisely if G is small. This is an intrinsic limitation of the model.

3 Estimation of the model

We will estimate the model by maximum likelihood. To do so we require the following assumptions.

Assumptions A. (i) The c_i^o are iid $N\left(0, \sigma_{c^o}^2\right)$. (ii) The w_g^o are iid $N\left(0, \sigma_{w^o}^2\right)$. (iii) The u_{it}^o are iid $N\left(0, \sigma_{u^o}^2\right)$. (iv) The c_i^* are iid $N^+\left(0, \sigma_{c^*}^2\right)$. (v) The w_g^* are iid $N^+\left(0, \sigma_{w^*}^2\right)$. (vi) The u_{it}^* are iid $N^+\left(0, \sigma_{u^*}^2\right)$. (Here N^+ denotes a half-normal distribution.)

B. $x_{it}, c_j^o, c_k^*, w_g^o, w_h^*, u_{ms}^o, u_{qr}^*$ are mutually independent for all $i, j, k, m, q = 1, \ldots, N$, $g, h = 1, \ldots, G$ and $r, s, t = 1, \ldots, T$.

These assumptions say that the regressors x_{ii} are exogenous in the strictest possible sense—independent of all observations on all of the error components—and can therefore be treated as fixed. The error components are independent across different components and different realizations of the same component. However, there is no restriction on the dependence of the x_{it} across firms or over time. These are strong assumptions, but they mirror the assumptions made in CKMV and Filippini and Greene for the model with two composed errors.

Define $\varepsilon_{it} = c_i + w_{g(i)} + u_{it}$. We have independence of the ε's across different groups, but within a group we have correlation across individuals and over time because of the group effect. Suppose that within group g we re-index the individuals as $i = 1, \ldots, n_g$ (a separate re-indexing for each group), and we let $\varepsilon_{(g)} = \left(\varepsilon_{11}, \ldots, \varepsilon_{1T}, \ldots, \varepsilon_{n_g 1}, \ldots, \varepsilon_{n_g T}\right)'$ be the $Tn_g \times 1$ vector of ε's for group g. Let $f_g\left(\varepsilon_{(g)}\right)$ be the density of $\varepsilon_{(g)}$. Define $y_{(g)}$and$x_{(g)}$ analogously to $\varepsilon_{(g)}$ and $d_{(g)} = (d_1, \ldots, d_T, \ldots, d_1, \ldots, d_T)'$,, so that $\varepsilon_{(g)} = y_{(g)} - x_{(g)}\beta - d_{(g)}\theta$. Then the log-likelihood function is

$$\ln L = \sum_{g=1}^{G} \ln f_g(y_{(g)} - x_{(g)}\beta - d_{(g)}\theta). \tag{2}$$

We will maximize this to calculate the maximum likelihood estimator. The parameters with respect to which the likelihood is maximized are $\beta, \theta, \sigma_{c^o}^2, \sigma_{c^*}^2, \sigma_{w^o}^2, \sigma_{w^*}^2, \sigma_{u^o}^2$ and $\sigma_{u^*}^2$.

The remaining issue is how to calculate the density f_g for each group g. The vector of random elements on which $\varepsilon_{(g)}$ depends is $\xi_g = (c_1, \ldots, c_{n_g}, w_g, u_{11}, \ldots, u_{n_g T})$, a vector of $Tn_g + n_g + 1$ independent skew-normal random variables. These have densities that depend on the univariate normal pdf and cdf; for example, $f_c(c_i) = \frac{2}{\sigma_c}\varphi\left(\frac{c_i}{\sigma_c}\right)\Phi\left(-\lambda_c \frac{c_i}{\sigma_c}\right)$, where $\sigma_c^2 = \sigma_{c^o}^2 + \sigma_{c^*}^2$ and $\lambda_c = \sigma_{c^*}/\sigma_{c^o}$, and similarly for w_g and u_{it}. Because the elements of ξ_g are mutually independent, the density of ξ_g is:

$$f_\xi\left(\xi_g\right) = f_g\left(w_g\right) \prod_{i=1}^{n_g} \left[f_c(c_i) \prod_{t=1}^{T} f_u(u_{it}) \right]. \tag{3}$$

Then we can obtain the density of $\varepsilon_{(g)}$ by integrating out w_g and the c_i:

$$f_g(\varepsilon_{(g)}) = \int \ldots \int f_g(w_g) \prod_{i=1}^{n_g} \left[f_c(c_i) \prod_{t=1}^{T} f_u(\varepsilon_{it} - w_g - c_i) \right] dw_g dc_1 \ldots dc_{n_g} \quad (4)$$

This integral could in principle be evaluated numerically, but it is of dimension $n_g + 1$ and a numerical evaluation would be slow and of questionable accuracy. We therefore seek alternatives to a brute-force numerical integration procedure. There are two such alternatives, which are extensions of procedures considered in the simpler four-component case.

The first alternative is to follow the path of CKMV (2014) and use results on the closed skew-normal family. The relevance of the closed skew-normal family to stochastic frontier models was pointed out by Domínguez-Molina et al. (2003) and González-Farías et al. (2004a, b). Skew-normal random variables like c_i, w_g and u_{it} are special cases of closed skew-normal random variables. Since independent closed skew-normal random variables are jointly closed skew-normal, ξ_g is closed skew-normal. Since linear combinations of closed skew-normal random variables are closed skew-normal, $\varepsilon_{(g)}$ is closed skew-normal. This makes it possible to write the density of $\varepsilon_{(g)}$ in an explicit compact form. However, it does not make it easy to calculate, because the explicit form involves evaluating the cdf of a multivariate normal distribution of dimension $n_g(T + 1)+1$. (In the simpler four-component model, this was a multivariate normal distribution of dimension $T + 1$.) See "Appendix" for some algebraic details.

A second alternative that is more numerically promising is to follow the logic of Filippini and Greene (2016) and calculate the likelihood by simulation. Conditional on $w_g, c_1, \ldots, c_{n_g}$, $(\varepsilon_{it} - w_g - c_i)$ for $i = 1, \ldots, n_g$, $t = 1, \ldots, T$ are i.i.d. with density f_u. Therefore, the density of $\varepsilon_{(g)}$ conditional on $w_g, c_1, \ldots, c_{n_g}$ is

$$f_{cond}(\varepsilon_{(g)}) = \prod_{i=1}^{n_g} \prod_{t=1}^{T} f_u(\varepsilon_{it} - w_g - c_i) \quad (5)$$

and

$$\begin{aligned} f_g(\varepsilon_{(g)}) &= \int \ldots \int f_{cond}(\varepsilon_{(g)}) f_g(w_g) \prod_{i=1}^{n_g} f_c(c_i) dw_g dc_1 \ldots dc_{n_g} \\ &= E f_{cond}(\varepsilon_g) \end{aligned} \quad (6)$$

where "E" represents the expectation over the distribution of $w_g, c_1, \ldots, c_{n_g}$. This expectation can be evaluated by averaging over simulated draws. Let $s = 1, \ldots, S$ index replications for the simulated draws, where S is a large number. For replication s, draw $w_g^{(s)}, c_1^{(s)}, \ldots \infty, c_{n_g}^{(s)}$ from the applicable skew-normal distributions (that is,

form them as the difference of a draw of a normal and a draw of a half-normal), and then average $f_{cond}(\varepsilon_{(g)})$ over these draws. Then our simulated density for group g is:

$$\hat{f}_g(\varepsilon_g) = \frac{1}{S}\sum_{s=1}^{S}\prod_{i=1}^{n_g}\prod_{t=1}^{T} f_u\left(\varepsilon_{it} - w_g^{(s)} - c_i^{(s)}\right). \tag{7}$$

Finally, then, the simulated log likelihood is $\ln \hat{L} = \sum_{g=1}^{G} \ln \hat{f}_g(y_{(g)} - x_{(g)}\beta - d_{(g)}\theta)$.

4 Prediction of the inefficiencies

Consider firm i at time t. It is a member of group $g(i)$, which for simplicity we will simply call group g. We wish to calculate the predicted inefficiencies $\hat{c}_i^*$, $\hat{w}_g^*$ and $\hat{u}_{it}^*$, which are the expectations of c_i^*, w_g^* and u_{it}^* conditional on $\varepsilon_{(g)} = \left(\varepsilon_{11}, \ldots, \varepsilon_{1T}, \ldots, \varepsilon_{n_g 1}, \ldots, \varepsilon_{n_g T}\right)'$. The reason that the conditioning set should be $\varepsilon_{(g)}$ is as follows. We have independence across groups, so values of ε_{jt} for firms j not in group g are irrelevant. However, we have correlation across firms in group g because of the common group effect w_g, so the values of ε_{jt} for all firms j that are in group g are relevant. That is why we need to evaluate

$$\hat{c}_i^* = E(c_i^* | \varepsilon_{(g)}) \tag{8}$$

and not the simpler expression

$$\tilde{c}_i^* = E(c_i^* | \varepsilon_{i1}, \ldots, \varepsilon_{iT}). \tag{9}$$

Assuming the conditional expectation functions in (8) and (9) to be known (more on that below), $\hat{c}_i^*$ is a more precise (smaller mean square error) predictor than $\tilde{c}_i^*$.

Using arguments similar to those in CKMV (2014), we could use the properties of the skew-normal distribution to derive explicit expressions for these conditional expectations. In CKMV, these expressions involved the cdf of a normal distribution of dimension $T + 1$, and its evaluation was feasible (though obviously this must depend on the value of T). Similarly, for $\tilde{c}_i^*$ as given in Eq. (9), we would need to evaluate the cdf of a normal distribution of dimension $T + 2$. However, for $\hat{c}_i^*$ as given in Eq. (8), we would need to evaluate the cdf of a normal distribution of dimension $n_g(T + 1) + 1$. For empirically reasonable values of n_g and T, this is unlikely to be feasible.

An alternative is to estimate the conditional expectation function in (8) nonparametrically. This is essentially the same strategy as in Amsler et al. (2014), in a different setting. We can consider using nearest neighbors or kernel nonparametric estimates.

To fix ideas, we will first give a brief summary of the nearest neighbors estimator in a generic setting. Suppose we have a scalar r (in our case, c_i^*) and a $k \times 1$ vector z (in our case, an estimate of ε_g). We want to estimate $\mu(z) = E(r|z)$ based on a sample $\{z_i, r_i, i = 1, \ldots, n\}$. Note that "$z$" is an arbitrary point, not necessarily one of the data points, and similarly "r" need not be observed. Let D(z, z^*) be a distance function

for k-dimensional vectors, generally of the form $\mathrm{D}(z, z^*) = (z^* - z)' A (z^* - z)$, where A is a positive definite matrix, such as $A = \left[\sum_{i=1}^{n} (z_i - \bar{z})(z_i - \bar{z})'\right]^{-1}$. Then $\hat{\mu}(z) = \left(\frac{1}{M}\right) \sum_j r_j$ where (i) M is an integer, and (ii) the sum is over the values of j such that z_j is one of the M nearest neighbors of z. That is, we pick the M values of j such that $\mathrm{D}(z, z_j)$ is smallest, and then we average the corresponding r_j values. This is pretty simple and it requires only the choice of M. For consistency of $\hat{\mu}(z)$ we require $M \to \infty$ and $M/n \to 0$ as $n \to \infty$. The point is that M needs to be big enough for our average to be meaningful, but small enough that the M nearest neighbors of z are quite close to z.

Coming back to our problem, "r" corresponds to c_i^* for some observation "i" in our real data sample; it is not observed, but we only want to estimate its conditional expectation. The first step is to estimate our model on the real data set. This yields estimates of the parameters, both the regression function parameters β and θ and the variance parameters in the distributions of c, w and u. Also then we can calculate $\hat{\varepsilon}_{it} = y_{it} - x_{it}' \hat{\beta} - d_t' \hat{\theta}$ and we can construct $\hat{\varepsilon}_{(g)} = \left(\hat{\varepsilon}_{11}, \ldots, \hat{\varepsilon}_{1T}, \ldots, \hat{\varepsilon}_{n_g 1}, \ldots, \hat{\varepsilon}_{n_g T}\right)$, which corresponds to "z" in the generic discussion above. Now we will construct a sample by simulation. Pick a very large value of S, the sample size in our simulated data set. For simulated observation $s = 1, \ldots, S$, generate

$$\varepsilon_{(g)}(s) = \left(\varepsilon_{11}(s), \ldots, \varepsilon_{1T}(s), \ldots, \varepsilon_{n_g 1}(s), \ldots, \varepsilon_{n_g T}(s)\right)' \tag{10}$$

from the appropriate closed skew-normal distribution. The easiest way to do this is to make random draws from the normal distributions of all of the c_i^o, w_g^o and u_{it}^o and from the half-normal distributions of the c_i^*, $w_g^*$and u_{it}^*, and then to construct $\varepsilon_{it}(s) = c_i^o(s) - c_i^*(s) + w_{g(i)}^o(s) - w_{g(i)}^*(s) + u_{it}^o(s) - u_{it}^*(s)$. Then use Eq. (10) to construct $\varepsilon_g(s)$. Note that $\varepsilon_g(s)$ corresponds to z_i in the generic discussion above, and $c_i^*(s)$ corresponds to r_i.

Finally, the nearest neighbors estimate of $\hat{c}_i^* = E(c_i^* | \varepsilon_{(g)})$ is the average of $c_i^*(s)$ over the M values of s such that $D\left(\varepsilon_{(g)}(s), \hat{\varepsilon}_{(g)}\right)$ is smallest, that is, of the M nearest neighbors of $\hat{\varepsilon}_{(g)}$ in the simulated data set.

The construction of the estimates of w_g^* or u_{it}^* is essentially the same. We simply average the values of $w_g^*(s)$ or $u_{it}^*(s)$ instead of $c_i^*(s)$. Also, the same procedure can be easily modified to estimate the simpler conditional expectation $\tilde{c}_i^*$ given in Eq. (9) above.

An alternative way to estimate the necessary conditional expectations is to use a kernel. We will briefly describe the well-known Nadaraya–Watson estimator in the same generic setting as above. (The logical step from the generic discussion to our specific case is exactly the same as for nearest neighbors.) Here we have $\hat{\mu}(z) = \sum_{j=1}^{n_g} w_j(z) r_j \Big/ \sum_{j=1}^{n_g} w_j(z)$, where w is a weighting function of the form

$$w_j(z) = K\left(\frac{z_j - z}{h}\right). \tag{11}$$

In this expression, K is a "kernel" function such that $K(z) \to 0$ as $|z| \to \infty$. For example, the Gaussian kernel is the standard multivariate normal pdf. And h is the

"bandwidth" which satisfies $h \to 0$ and $nh \to \infty$ as $n \to \infty$. In nearest neighbors we have weights that are either zero or one, whereas in kernel regression every observation gets positive weight but how much depends on how close the particular z_j is to z.

There is a vast literature on the choice of kernel and bandwidth. See, for example, Li and Racine (2006); or, in this day and age, Google.

We now return to a point we made earlier, concerning the comparison of $\hat{c}_i^*$ and $\tilde{c}_i^*$ as given in Eqs. (8) and (9) above. If these conditional expectations are known, then $\hat{c}_i^*$ is a more precise (smaller mean square error) prediction of c_i^* than $\tilde{c}_i^*$, because it has a larger conditioning set. However, if they are estimated nonparametrically, the so-called curse of dimensionality applies, and the estimation error in the evaluation of $\hat{c}_i^*$ is larger than the estimation error in the evaluation of $\tilde{c}_i^*$. This may not matter much, because the nonparametric estimation error goes away asymptotically, as $S \to \infty$, where S is the number of simulation draws, and we can make that as large as we want. The only constraint is computing time.

5 Concluding remarks

The aim of this paper was to provide a stochastic frontier model that captures accurately the metafrontier concept. Firms are members of groups, and we model the inefficiency of a given firm relative to its group-specific frontier, and we also model the inefficiency of a given group relative to the overall metafrontier.

In our model the parameters of the deterministic portion of the frontier (β and θ) are the same for all groups. This assumption could be relaxed if there are enough firms per group.

While the model is (in our opinion) attractive, further research will be needed to see how useful it is empirically. Here we see two main issues. The first issue is that the methods we propose for estimation of the parameters and for the prediction of the inefficiencies are computationally intensive. This is probably not an insurmountable obstacle. The second issue is that the decomposition of the group effects into their heterogeneity and inefficiency components will require estimates of the variances of w^o and w^*, and these parameters cannot be expected to be estimated very precisely if the number of groups (G) is small. This is an intrinsic limitation of the model.

How serious this second issue is clearly depends on the nature of the data and the application. For many production metafrontiers problems, the number of groups will be small. However, stochastic frontier models can be applied in other settings, and in some of them the number of groups may not be small. For example, suppose that an educational researcher is interested in the efficiency of schools and school districts in "producing" a measurable outcome, like the average student score on a standardized test. Then we would likely have a small number of time periods (T), a small number of schools per district (n_g), but a large number of districts (G).

As noted in the Introduction, it is assumed that the number of groups and the group membership of each firm are known. This is true throughout the metafrontier literature. An interesting and non-trivial question is whether this assumption could be relaxed. In principle this is probably possible; we don't see an identification problem here. However, we do not see any straightforward or attractive way to relax the assumption

that the number of groups is known. Given a known number of groups, the assumption that the group memberships are known could probably be relaxed. For example, we could propose a "latent class" model in which there are probabilities for firms to be in each of the G groups, perhaps depending on firm characteristics and some auxiliary parameters, and then these probabilities would be estimated along with the production function parameters. As is standard in the latent class literature, these would be the unconditional probabilities of group membership, and we could also calculate the probabilities conditional on the data. This kind of extension of our model is intuitively reasonable but clearly beyond the scope of this paper.

Compliance with ethical standards

Conflict of interest All of the authors declare that they have no conflicts of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Appendix

We wish to derive an expression for the density of $\varepsilon_{(g)} = \left(\varepsilon_{11}, \ldots, \varepsilon_{1T}, \ldots, \varepsilon_{n_g 1}, \ldots, \varepsilon_{n_g T}\right)'$. The starting point will be the density of $\xi_g = \left(c_1, \ldots, c_{n_g}, w_g, u_{11}, \ldots, u_{n_g T}\right)$. Since we are dealing with a specific group, group g, we will simplify the notation, for this "Appendix" only, by omitting the subscript g. Thus we write ε in place of ε_g, ξ in place of ξ_g, w in place of w_g and n in place of n_g. We will use results on the closed skew-normal distribution from González-Farías et al. (2004b) (hereafter GDG).

A p-dimensional random variable Z is distributed as $\mathrm{CSN}_{p,q}(\mu, \Sigma, D, \nu, \Delta)$ if its density is $f(z) = C\varphi_p(z; \mu, \Sigma)\Phi_q(D(z-\mu); \nu, \Delta)$. Here φ_p and Φ_q are the p-variate normal density and the q-variate normal cdf, respectively, and $C^{-1} = \Phi_q\left(0; \nu, \Delta + D\Sigma D'\right)$. The dimensions of the parameters are as follows: $\mu : p \times 1$, $\Sigma : p \times p$, $D : q \times p$, $\nu : q \times 1$, $\Delta : q \times q$. The relevance of this to the our model is that the composed error c_i, with parameters λ_c and σ_c^2, is distributed as $\mathrm{CSN}_{1,1}\left(0, \sigma_c^2, -\frac{\lambda_c}{\sigma_c}, 0, 1\right)$, and similarly for w and u_{it}.

Proposition 2.4.1 of GDG says that independent marginally CSN random variables are jointly CSN. Generically, if $Z = \left(Z_1', \ldots, Z_k'\right)'$ where the Z_j are mutually independent and $Z_j \sim \mathrm{CSN}_{p_j,q_j}\left(\mu_j, \Sigma_j, D_j, \nu_j, \Delta_j\right)$, then $Z \sim CSN_{p,q}(\mu, \Sigma, D, \nu, \Delta)$, where $p = \sum_j p_j$, $q = \sum_j q_j$, $\mu = \left(\mu_1', \ldots, \mu_k'\right)'$, $\nu = \left(\nu_1', \ldots, \nu_k'\right)'$, $\Sigma = \oplus_{j=1}^k \Sigma_j$, $D = \oplus_{j=1}^k D_j$, $\Delta = \oplus_{j=1}^k \Delta_j$. Here $\oplus$ is the matrix direct sum operator that makes matrices A and B into a block diagonal matrix: $A \oplus B = \begin{bmatrix} A & O \\ O & B \end{bmatrix}$. In our case, this implies that $\xi \sim \mathrm{CSN}_{q,q}(\mu, \Sigma, D, \nu, \Delta)$ where $q = n(T+1)+1$ and:

$$\mu = 0, \nu = 0 \text{ (both } q \times 1)$$

$$\Sigma = \sigma_c^2 I_n \oplus \sigma_w^2 \oplus \sigma_u^2 I_{nT} \quad \text{(a diagonal matrix of dimension } q\text{)}$$
$$D = -\frac{\lambda_c}{\sigma_c} I_n \oplus -\frac{\lambda_w}{\sigma_w} \oplus -\frac{\lambda_u}{\sigma_u} I_{nT} \quad \text{(a diagonal matrix of dimension } q\text{)}$$
$$\Delta = I_q$$

Proposition 2.3.1 of GDG says that linear combinations of jointly CSN random variables are jointly CSN. Generically, suppose that $Z \sim \text{CSN}_{p,q}(\mu, \Sigma, D, \nu, \Delta)$ and let A be $m \times p$, $m \leq p$, $\text{rank}(A) = m$. Then $AZ \sim \text{CSN}_{m,q}(\mu_A, \Sigma_A, D_A, \nu, \Delta_A)$, where $\mu_A = A\mu$, $\Sigma_A = A\Sigma A'$, $D_A = D\Sigma A' \Sigma_A^{-1}$, $\Delta_A = \Delta + D\Sigma D' - D\Sigma A' \Sigma_A^{-1} A \Sigma D'$. In our case, $\varepsilon = A\xi$ where A is $nT \times q$ and is defined as follows:

$$A = [B_1, B_2, B_3]$$
$$B_1 = I_n \otimes 1_T, \text{ where } 1_T \text{ is a } T \times 1 \text{ vector of ones(so } B_1 \text{ is of dimension } nT \times n)$$
$$B_2 = 1_{nT} \ (nT \times 1)$$
$$B_3 = I_{nT} \ (nT \times nT)$$

Therefore, $\varepsilon \sim CSN_{nT,q}(\mu_A, \Sigma_A, D_A, \nu, \Delta_A)$ and the density of ε is

$$f(\varepsilon) = C_A \varphi_{nT}(\varepsilon; \mu_A, \Sigma_A) \Phi_q(D_A(\varepsilon - \mu_A); \nu, \Delta_A) \tag{A1}$$

where $C_A^{-1} = \Phi_q\left(0; \nu, \Delta_A + D_A \Sigma_A D_A'\right)$.

Some of these arguments of the density can be simplified. For example, $\mu_A = 0$, $\nu = 0$, and $\Sigma_A = \sigma_u^2 I_{nT} + \sigma_c^2 \left(I_n \otimes 1_T 1_T^{'}\right) + \sigma_w^2 1_{nT} 1_{nT}'$. However, D_A and Δ_A are rather complicated, and, importantly, Δ_A does not have any special algebraic structure (e.g., block diagonal) that would allow the dimensionality of the integral implicit in the cdf $\Phi_q(D_A \varepsilon; 0, \Delta_A)$ to be reduced. So we are left with the task of evaluating the joint cdf of a multivariate normal of dimension $q = n(T + 1) + 1$. This is not likely to be practical.

References

Amsler C, O'Donnell CJ, Schmidt P. Stochastic metafrontiers. Econom Rev. 2017;36:1007–20.

Amsler C, Prokhorov A, Schmidt P. Using copulas to model time dependence in stochastic frontier models. Econom Rev. 2014;33:497–522.

Azzalini A. A class of distributions which includes the normal ones. Scand J Stat. 1985;12:171–8.

Battese GE, Rao DSP. Technology gap, efficiency, and a stochastic metafrontier function. Int J Bus Econ. 2002;1:87–93.

Battese GE, Rao DSP, O'Donnell CJ. A metafrontier production function for estimation of technical efficiencies and technology gaps for firms operating under different technologies. J Prod Anal. 2004;21:91–103.

Colombi R, Martini G, Vittadini G (2011) A stochastic frontier model with short-run and long-run inefficiency random effects. Working Paper, University of Bergamo.

Colombi R, Kumbhakar SC, Martini G, Vittadini G. Closed skew normality in stochastic frontiers with individual effects and long/short-run inefficiency. J Prod Anal. 2014;42:123–36.

Domínguez-Molina JA, González-Farías G, Ramos-Quiroga R (2003) Skew normality in stochastic frontier analysis. Commun Tech. I-03-18, 1–13.

Filippini M, Greene W. Persistent and transient productive inefficiency: a maximum simulated likelihood approach. J Prod Anal. 2016;45:187–96.
Fuller WA, Battese GE. Transformations for estimation of linear models with nested-error structure. J Am Stat Assoc. 1973;68:626–32.
González-Farías G, Domínguez-Molina JA, Gupta AK. Additive properties of skew normal random vectors. J Stat Plan Inference. 2004a;126:521–34.
González-Farías G, Domínguez-Molina JA, Gupta AK. The closed skew normal distribution. In: Genton M, editor. Skew elliptical distributions and their applications: a journey beyond normality. Boca Raton: Chapman and Hall; 2004b.
Hayami Y, Ruttan VW. Agricultural development: an international perspective. Baltimore: The Johns Hopkins University Press; 1971.
Hayami Y, Ruttan VW. Agricultural development: an international perspective. Revised and expanded ed. Baltimore: The Johns Hopkins University Press; 1985.
Kim J-S, Frees EW. Multilevel modelling with correlated effects. Psychometrika. 2007;72:505–33.
Kumbhakar SC, Lien G, Hardaker JB. Technical efficiency in competing panel data models. J Prod Anal. 2014;41:321–37.
Lai H-P, Kumbhakar SC. Panel data stochstic frontier model with determinants of persistent and transient inefficiency. Eur J Oper Res. 2018;271:746–55.
Lau LJ, Yotopoulos PA. The meta-production function approach to technological change in world agriculture. J Dev Econ. 1989;31:241–69.
Li Q, Racine J. Nonparametric econometrics: theory and practice. London: Princeton University Press; 2006.
Matyas L, editor. The econometrics of multi-dimensional panels. Berlin: Springer; 2017.
Moreira V, Bravo-Ureta B. Technical efficiency and metatechnology ratios for dairy farms in three southern cone countries: a stochastic meta-frontier model. J Prod Anal. 2010;33:33–45.
O'Donnell CJ, Rao DSP, Battese GE. Metafrontier frameworks for the study of firm-level efficiencies and technology ratios. Empir Econom. 2008;34:231–55.
Pitt MM. Farm-level fertilizer demand in Java: a meta-production function approach. Am J Agric Econ. 1983;65:502–8.
Raudenbush SW, Bryk AS. Hierarchical linear models: applications and data analysis methods. 2nd ed. London: Sage Publications; 2002.
Tsionas EG, Kumbhakar SC. Firm heterogeneity, persistent and transient technical inefficiency: a generalized true random effects model. J Appl Econom. 2014;29:110–32.
Villano R, Bravo-Ureta B, Solis D, Fleming E. Modern rice technologies and productivity in the Philippines: disentangling technology from managerial gaps. J Agric Econ. 2015;66:129–54.
Wooldridge JM. Econometric analysis of cross section and panel data. 2nd ed. Cambridge: MIT Press; 2010.
Yang Y, Schmidt P (2020) An econometric approach to the estimation of multi-level models. J Econom. (forthcoming).